Diminishing Returns as a Lever for Fair Cooperation in Multi-Agent Reinforcement Learning

Claude Formanek[1 [0000−0003−4738−2088]], Karabo Letsholo[1 [0009−0002−1389−6355]], and Jonathan P. Shock[1 [0000−0003−3757−0376]]

University of Cape Town

Abstract. We study reward design for prosocial behaviour in multi-agent reinforcement learning (MARL) under a social-dilemma setting. In an apple-harvesting environment where agents must both harvest apples and clean a river to sustain yields, we compare four schemes: independent rewards, fully shared team rewards, fractional team rewards, and an inequality-aware diminishing-returns shaping. Independent rewards induce over-harvesting and under-provision of cleaning, collapsing group productivity. Shared rewards achieve high total output via division of labour but produce large disparities in individual payoffs. Fractional team rewards fail to correct incentives: because self-harvesting strictly dominates spillovers from teammates, agents continue to prioritise harvesting over cleaning. In contrast, the inequality-aware shaping performs strongly: agents harvest greedily at first, then cross a utility threshold and reallocate effort to cleaning, which raises group productivity and the marginal scope for individual gains. This scheme nearly matches the shared-reward baseline in total apples while maintaining low inequality. Our contributions are: (i) an empirical demonstration of the efficiency–equity trade-off inherent in standard reward extremes, (ii) framing inequality as a controllable dimension of MARL objectives, and (iii) practical reward designs that encourage reciprocity and public-good provision. We discuss implications for training prosocial agents in larger-scale systems, including language agents, robots, and mixed human–AI teams.

Keywords: Multi-Agent Reinforcement Learning · Social Dilemmas · Reward Shaping · Prosocial Behavior · Cooperation · Inequality · Fairness

1 Introduction

Reinforcement learning (RL) has emerged as a powerful paradigm for training agents to make sequential decisions in complex environments (42). After a period of doubt in the utility of the field to real-world settings (11), interest in RL has resurged, driven in part by its central role in aligning large language models (LLM) with human feedback (3) and by advances in high-fidelity simulators that enable scalable data generation in robotics and related fields (28). As AI systems

become increasingly agentic (i.e. able to act autonomously to achieve goals), they will more frequently encounter and interact with other agents. In such multi-agent settings, coordination and cooperation become essential concerns (18). If future AI systems are to function harmoniously alongside one another and with humans, they must not only be capable of optimising individual objectives but also of exhibiting prosocial behaviour: seeking win–win outcomes whenever possible and being aware of the social consequences of their actions.

Multi-agent reinforcement learning (MARL) (1) naturally gives rise to tensions between individual and collective incentives, especially in social dilemmas (26) exemplified by the "tragedy of the commons" (34; 35). Conventional reward engineering in MARL sits at two extremes. On one end, fully independent rewards (each agent is rewarded solely for its own achievements), which often lead to short-sighted behaviour that is individually rational yet collectively harmful. On the other end, fully cooperative team rewards (every agent receives the same payoff based on team performance), which can improve collective outcomes but conflate credit assignment (33) and may conceal or even exacerbate inequities in contribution and payoff across agents (22). In contexts where fairness and equality matter, large disparities in individual returns can be undesirable, even if the aggregate outcome is high (50). This raises a central challenge: how can we design reward schemes that maximise efficiency gains while mitigating inequality among agents?

We investigate this question in a stylised but illustrative MARL environment: an apple-harvesting game called CleanUp in which a river must be kept clean for the orchard to remain productive (22). Agents can either be harvesting apples or cleaning the river; doing both tasks is necessary for sustained success. This setting manifests a classic social dilemma: harvesting yields immediate private benefit, while cleaning creates a positive externality that enables future harvests for all. Our initial experiments mirror intuition. Under independent rewards for picking apples, agents over-exploit the environment and under-provide cleaning, leading to a polluted river and a collapse in team yield. Under a fully shared team reward, the group achieves markedly higher overall return, typically by settling into a division of labour where some agents specialise in harvesting while others specialise in cleaning. However, this cooperative solution frequently produces substantial inequality in per-agent returns, with cleaners receiving much less reward than harvesters despite their essential contribution to team success.

To reconcile this tension between efficiency and equity, we study two reward shaping techniques intended to promote prosocial behaviour without sacrificing collective performance. First, we propose *fractional team rewards*: in addition to the primary reward an agent receives when it personally harvests an apple, each agent receives a smaller "fractional" reward whenever any teammate harvests. Intuitively, this scheme partially integrates teammates' benefits, encouraging reciprocity and coordination while preserving incentives for individual effort. Second, we explore an inequality-aware shaping scheme inspired by *diminishing marginal utility* (29). Here, the marginal value of additional apples decreases as an agent's cumulative harvest grows (especially relative to its peers) so that

agents who are already far ahead discount further gains and are nudged toward contributing to public goods (e.g., cleaning), while agents who are behind value marginal harvests more. Both designs aim to maintain high team output while reducing disparities in individual outcomes.

We present a pilot empirical study comparing independent rewards, fully shared team rewards, fractional team rewards, and inequality-aware diminishing returns in the CleanUp environment. We evaluate each scheme along two axes: group productivity (measured by the total number of apples collected) and the inequality of outcomes across agents. Consistent with our intuition, independent rewards yield poor collective performance; fully shared rewards substantially improve total yield but often exhibit high inequality at the agent level. Furthermore, our experiments show that fractional team rewards do not remedy this, because an apple collected by oneself still dominates the smaller reward from a teammate's harvest, and agents therefore continued to prefer self-harvesting over cleaning. By contrast, the inequality-aware diminishing returns performed very well. Early on in training, agents began by harvesting greedily, but once their personal return crossed a certain threshold, they *suddenly* shifted effort toward cleaning, which raised team productivity and, in turn, the scope for individual gains. This scheme nearly matches the shared reward in total output while keeping inequality very low. Together, these results suggest a promising path toward training prosocial agents that are both collectively effective and individually fairer.

Our contributions are threefold. First, we provide a clear empirical demonstration of how standard reward extremes (fully independent and fully cooperative) trade off efficiency and equity in a concrete social dilemma. Second, we highlight inequality as a salient and measurable dimension of multi-agent outcomes, complementary to aggregate performance, and show that it can be meaningfully influenced through reward shaping. Third, we introduce and evaluate two practical reward schemes—fractional team rewards and inequality-aware diminishing returns—that promote coordination and reciprocity while mitigating disparities in payoffs. While our study is deliberately simple, we believe it uncovers design principles that can inform the training of larger-scale agentic systems, spanning LLM agents (19), robots (44), and mixed human–AI systems like self-driving cars (7).

The remainder of the paper is organised as follows. We first review background on RL and MARL, together with relevant notions from game theory and economics related to social welfare and inequality. We then discuss related work on prosocial RL, reward shaping, and cooperation in social dilemmas. Next, we present our experimental environment and methods, detailing the two proposed reward designs and how they are instantiated. We report experimental results and analyses, before finally concluding by summarising key findings, limitations, and directions for future work.

2 Background

Reinforcement learning. We model sequential decision making as a Markov decision process (MDP) (21) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, defined as a tuple of states $\mathcal{S}$, actions, $\mathcal{A}$, transition dynamics given by $\mathcal{P}(s' \mid s, a)$, rewards $\mathcal{R}$, and discount factor $\gamma \in (0, 1]$, with trajectories of states, actions and rewards $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$ induced by policy $\pi_\theta(a \mid s)$. The goal of an RL agent is to maximise the discounted return

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\Big[\sum_{t=0}^{T-1} \gamma^t \, r(s_t, a_t)\Big],$$

with horizon T (finite or infinite). Defining the value functions (4)

$$V^\pi(s) = \mathbb{E}[\sum_{t \geq 0} \gamma^t r_t \mid s_0{=}s]\,,$$
$$Q^\pi(s, a) = \mathbb{E}[\sum_{t \geq 0} \gamma^t r_t \mid s_0{=}s, a_0{=}a]\,,$$

and the advantage

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)\,,$$

The policy gradient theorem (43) gives the gradient of the objective function in terms of the gradient of the log of the policy distribution:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s \sim d^{\pi_\theta},\, a \sim \pi_\theta}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, A^{\pi_\theta}(s, a)\big],$$

where d^{π_θ} is the discounted state visitation distribution and θ parameterises the policy, generally via a neural network (NN). In practice, one uses an estimator $\widehat{A}_t$ (e.g., generalised advantage estimation (38)) and a learned baseline to reduce variance.

In this work, we will utilise Proximal Policy Optimisation (PPO) (39), which has become a state-of-the-art RL algorithm across many benchmarks. PPO stabilises on-policy updates by constraining policy changes via a clipped surrogate objective. The policy update magnitude is defined in terms of the importance ratio

$$r_t(\theta) = \tfrac{\pi_\theta(a_t \mid o_t)}{\pi_{\theta_{\text{old}}}(a_t \mid o_t)}\,,$$

and the clipped objective is then given in terms of the advantage estimator and the change in the policy, and a parameter ϵ which controls the update trust region:

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}_t\Big[\min\big(r_t(\theta)\,\widehat{A}_t,\ \text{clip}(r_t(\theta), 1{-}\epsilon, 1{+}\epsilon)\,\widehat{A}_t\big)\Big]\,.$$

The PPO policy gradient then updates the policy parameters as

$$\nabla_\theta J_{\text{PPO}}(\theta) = \nabla_\theta\, \mathcal{L}_{\text{clip}}(\theta) + c_{\text{ent}}\, \mathbb{E}_t\big[\nabla_\theta\, \mathcal{H}(\pi_\theta(\cdot \mid o_t))\big].$$

Here, $\mathcal{H}(\pi_\theta(\cdot \mid o_t))$ is the entropy of the policy's action distribution at observation o_t, which for discrete actions is

$$\mathcal{H}(\pi_\theta(\cdot \mid o_t)) = -\sum_{a \in \mathcal{A}} \pi_\theta(a \mid o_t) \log \pi_\theta(a \mid o_t).$$

The coefficient $c_{\text{ent}}>0$ scales the entropy bonus to encourage exploration by preventing premature collapse to low-entropy (overly deterministic) policies; it is commonly annealed over training.

Multi-agent reinforcement learning. MARL is often formulated as a partially observable stochastic game (POSG) (20) $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}, P, \{r_i\}, \{\mathcal{O}_i\}, \gamma \rangle$, with agents $i \in \mathcal{I}=\{1, \ldots, N\}$. While the state space $\mathcal{S}$ remains the same as in the single agent case, in contrast to the single agent, fully observable case, each agent, i, may only have access to a set of observations $o_{i,t} \in \mathcal{O}_i$ and is able to perform a set of actions $\mathcal{A}_i$. Each agent has policy $\pi_{\theta_i}(a_i \mid o_i)$, where each agent may have its own parameterisation θ_i of its policy; the joint policy factorizes as $\pi_{\boldsymbol{\theta}}(\mathbf{a} \mid \mathbf{o}) = \prod_{i=1}^{N} \pi_{\theta_i}(a_i \mid o_i)$. The joint action/observation are $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$, $\mathbf{o}_t = (o_{1,t}, \ldots, o_{N,t})$; $P(s_{t+1} \mid s_t, \mathbf{a}_t)$ is the state transition kernel; the observation kernel $\Omega(\mathbf{o}_{t+1} \mid s_{t+1}, \mathbf{a}_t)$ generates next observations; $r_i(s_t, \mathbf{a}_t)$ gives agent i's reward (so $r_{i,t}=r_i(s_t, \mathbf{a}_t)$). The initial state distribution may also be provided with $s_0 \sim \rho_0$.

Depending on the structure of the reward functions, the game can be zero-sum (competitive), fully cooperative, or mixed-motive. Specifying a solution concept in MARL is more challenging than in single-agent RL, where the goal is simply to maximise return. In MARL, solution concepts are often borrowed from Game Theory, such as Nash equilibrium, Pareto optimality, and best-response dynamics (31). Thus, to fully specify a MARL problem, one needs to define both the POSG together with an appropriate solution concept (1).

A central challenge in MARL is non-stationarity (6). Standard RL assumes a stationary environment with fixed transition and reward functions. In MARL, however, each agent's policy updates change the effective dynamics faced by others, making the learning problem much harder than in the single-agent case. This undermines replay-based off-policy methods (12), since replayed experience quickly becomes stale. In contrast, on-policy algorithms (e.g., Independent PPO (8)) are typically more robust because they update from recent, in-distribution data.

For independent PPO, each agent maximizes its own objective

$$J_i(\theta_i) = \mathbb{E}[\sum_t \gamma^t r_{i,t}]\,,$$

with per-agent advantage $\widehat{A}_{i,t}$. The PPO importance ratio for agent i is $r_{i,t}(\theta_i) = \frac{\pi_{\theta_i}(a_{i,t}|o_{i,t})}{\pi_{\theta_i^{\text{old}}}(a_{i,t}|o_{i,t})}$. The per-agent clipped surrogate is then defined equivalently to the single-agent case, but not separately for each agent:

$$\mathcal{L}_{\text{clip},i}(\theta_i) = \mathbb{E}_t\Big[\min\big(r_{i,t}(\theta_i)\,\widehat{A}_{i,t},\ \text{clip}(r_{i,t}(\theta_i), 1-\epsilon, 1+\epsilon)\,\widehat{A}_{i,t}\big)\Big]\,.$$

The corresponding policy gradient for agent i therefore also mirrors the single-agent case:

$$\nabla_{\theta_i} J_{\mathrm{PPO},i}(\theta_i) = \nabla_{\theta_i} \mathcal{L}_{\mathrm{clip},i}(\theta_i) + c_{\mathrm{ent}}\, \mathbb{E}_t\big[\nabla_{\theta_i} \mathcal{H}(\pi_{\theta_i}(\cdot \mid o_{i,t}))\big],$$

where there will be a separate critic network for each agent with value-function parameters (e.g., ϕ_i) updated by a separate squared-error loss; optionally, a centralised critic can be used to compute $\widehat{A}_{i,t}$ with global information while retaining decentralised execution (48).

Payoffs, welfare, and inequality. Each agent i accrues an individual payoff equal to its discounted return $G_i = \mathbb{E}[\sum_{t=0}^{T-1} \gamma^t r_{i,t}]$; we denote the payoff profile by $\mathbf{G} = (G_1, \ldots, G_N)$. A strategic interaction is a social dilemma when individually rational (self-interested) behaviour leads to outcomes that are Pareto-inferior to more cooperative profiles (any change by any agent individually will lead to a worse outcome for that agent), typically due to negative externalities and under-provision of public goods (e.g., the tragedy of the commons).

There are different ways to define the welfare outcome (30; 5) of a given set of policies. Utilitarian welfare is an aggregate of the total performance, defined here as the average over discounted returns:

$$W_{\mathrm{util}} = \frac{1}{N} \sum_{i=1}^{N} G_i.$$

On the other hand, egalitarian welfare emphasises the worst-off agent (Rawlsian max–min) and is defined as:

$$W_{\mathrm{egal}} = \min_{i \in \{1,\ldots,N\}} G_i.$$

These capture complementary desiderata: efficiency (high total output) versus equity (no agent left behind).

Inequality, on the other hand, is often measured in economics by the Gini coefficient (40), which summarises the dispersion of outcomes. One definition uses pairwise differences:

$$\mathrm{Gini}(G_1, \ldots, G_N) = \frac{1}{2N^2\, W_{\mathrm{util}}} \sum_{i=1}^{N} \sum_{j=1}^{N} \big|G_i - G_j\big|.$$

An equivalent computational form (with $G_{(1)} \le \cdots \le G_{(N)}$ ordered) is

$$\mathrm{Gini} = \frac{1}{N\, W_{\mathrm{util}}} \sum_{i=1}^{N} (2i - N - 1)\, G_{(i)}.$$

It is easier to see from the first formulation that the Gini coefficient is 0 under perfect equality (all G_i=W_{util}), and from the second, it can be seen that it approaches 1 when a single agent captures nearly all returns.

3 Related Work

Cooperation in multi-agent systems is often studied through *sequential social dilemmas* in Markov games, where individually rational choices undermine long-run collective welfare. Canonical gridworld environments such as CleanUp capture this tension. Research on promoting cooperation and fairness in MARL has taken on a number of different directions.

Reward shaping for cooperation in sequential social dilemmas . In (22) Hughes et al. extended inequity aversion from static games to Markov games (from stateless to stateful), reshaping returns to reduce advantageous and disadvantageous inequality and improving cooperation in intertemporal social dilemmas. This reward shaping helps with temporal credit assignment and supports punishing defectors. Lupu and Precup (27) introduced "gifting" – allowing agents to transfer rewards to other agents – which stabilises prosocial behaviour and mitigates early-stage tragedies of the commons. Mannion et al. (10) showed that Potential-Based Reward Shaping (PBRS) in commons dilemmas (the *Tragic Commons* domain) can guide self-interested agents to conserve shared resources and achieve collective gains without altering the Nash equilibria. Intrinsic rewards have also been studied as a mechanism for influencing multi-agent behaviour (23).

Influence, reciprocity, and opponent shaping . Opponent-shaping methods like LOLA (13) influence co-learners by anticipating their updates but require heavy computation or unrealistic access. Recent work on reciprocal reward influence (51) introduces intrinsic objectives for reciprocation without differentiating thro-ugh opponents, improving cooperation in matrix and sequential games.

Fairness via welfare optimisation in MARL . Zimmer et al. (52) formulated fairness as optimising differentiable social welfare functions (e.g., generalised Gini, lexicographic maximin, proportional fairness) in decentralised cooperative MARL, using specialised architectures and policy-gradient methods. Jiang and Lu (24) proposed FEN, a hierarchical decentralised framework optimising both fairness and efficiency via sub-policy coordination. Domain-specific studies, such as Aloor et al. (2) and Grupen et al. (15), report that fairness can be improved with modest efficiency costs by incorporating fairness terms into reward functions and using equivariant policies.

Online fair allocation and Nash social welfare . Beyond control tasks, online decision-making with fairness objectives studies regret bounds for criteria like Nash social welfare (49). Results show sharp regret rates in stochastic settings, impossibility under adversarial bandit feedback, and restored tractability with full-information feedback. These findings formalise the difficulty of optimising equitable welfare online.

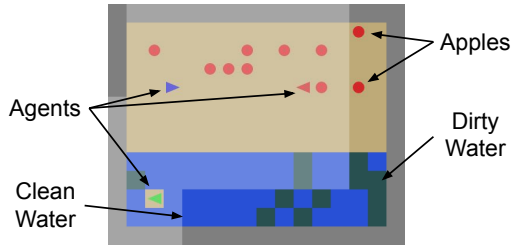


Fig. 1: CleanUp environment implemented in SocialJax (16). Grey shaded region represents the agents limited field of view.

Dual-reward MARL . DE-MADDPG (41) and similar approaches separate global and local critics to stabilise learning when optimising combined team and individual rewards, avoiding instability from entangled reward functions. Our diminishing-returns utility can be seen as a single-signal alternative, implicitly pricing the externality of being far ahead without requiring an explicit team-reward channel.

Credit assignment and surplus sharing . Difference rewards (9; 46) are an established credit assignment mechanism that quantifies individual contributions while preserving desirable equilibrium properties. Shapley value-based credit assignment (45) has also been proposed to fairly distribute surplus in cooperative MARL, improving stability and fairness.

4 Method

4.1 Experimental environment

We use the CleanUp environment from SocialJAX [1] (16). In CleanUp, agents collect apples for +1 reward/apple while keeping a shared river clean. Pollution builds up over time. Agents can use a *clean* action to remove some pollution. If pollution is high, apples stop regrowing; when the river is kept clean, apples regrow faster. Thus, taking apples without cleaning hurts future yield, while regular cleaning supports long-term rewards. Agents have a limited, egocentric field of view around themselves, showing local tiles (free space, walls), apples, river tiles with pollution level, and nearby agents. We process this local view with a small convolutional encoder, which feeds into the policy network. Actions for each agent are: move up, down, left, right; pick (harvest an adjacent apple); and clean (reduce pollution on an adjacent river tile). Cleaning has no immediate reward but improves future regrowth; picking gives immediate reward

[1] https://sites.google.com/view/socialjax/home

but can reduce future yield if overused. In our experiments, we configured the environment to have 3 agents because this was computationally tractable for us.

SocialJAX is implemented in JAX, which compiles and batches environment steps so one can run many copies in parallel on a single GPU/TPU. This is much faster than CPU-based suites like Melting Pot (reported speedups of around 50–140x for CleanUp), making experiments practical for limited compute. Even so, full training still takes up to 4 hours on one GPU, indicating that the environment remains challenging.

4.2 Algorithm and Model

We build on Independent PPO (IPPO) from JAXMARL (37) with parameter sharing (17) across agents. Each agent acts from its local view, which we encode with a small CNN (25) feeding two MLP (36) heads for policy and value.

We collect T-step rollouts from E parallel environments on a single GPU and run PPO with a clipped objective, GAE(λ), entropy bonus, minibatch SGD, advantage normalisation, gradient clipping, and linear learning-rate decay. The full pipeline—environment stepping, model forward, and updates—is written in JAX, compiled with `jit`, and vectorised with `vmap`/`scan`, enabling fast end-to-end training as in JAXMARL. All the code for this work will be fully open-sourced upon publication.

4.3 Reward shaping

In order to address the limitations of the fully independent and fully cooperative reward functions, we considered two reward shaping approaches (32; 47).

Fractional team reward: The most naive extension of the two extreme cases is a reward function that is a mix of the two. Intuitively, agents should be rewarded for picking up apples themselves, but also receive a smaller reward for when others pick up apples. One would imagine that this would lead to a more balanced behaviour, where agents are not only looking out for themselves, but also for others.

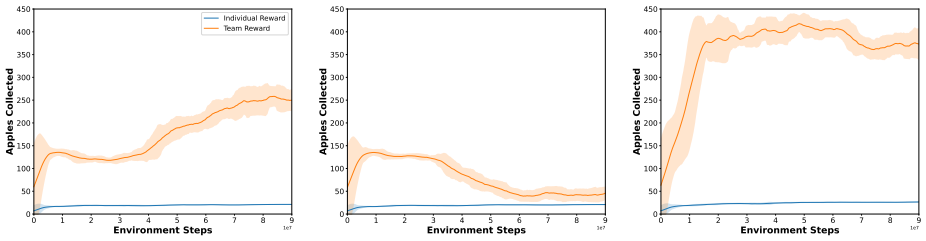
We define the reward function as follows:

$$r_i = r_i^{\text{ind}} + r_i^{\text{cop}}$$

where r_i^{ind} is the independent reward function and r_i^{cop} is the cooperative reward function. The independent reward function for each agent is simply a reward of $+1$ for picking up an apple and 0 otherwise. The cooperative reward function is defined as

$$r_i^{\text{cop}} = \alpha \cdot \sum_{j \in -i} r_j,$$

where $-i$ is the set of all agents excluding i, and α controls the magnitude of the cooperative reward. Typically, we set $\alpha = 0.1$, which means that agents receive a reward of 0.1 for each apple picked up by other agents.



(a) Agent 0 (b) Agent 1 (c) Agent 2

Fig. 2: Individual vs Team Rewards. Total apples collected by each agent per episode over the course of training.

Inequality-aware diminishing returns: In the second approach, we scale an agent's apple reward by how different its returns are from the agent with the highest return so far. Let $c_i(t)$ be the cumulative apples collected by agent i up to time t, and let $c_{\max}(t) = \max_j c_j(t)$. Define the normalised count

$$\tilde{c}_i(t) = \begin{cases} \dfrac{c_i(t)}{c_{\max}(t)} & \text{if } c_{\max}(t) > 0, \\ 0 & \text{if } c_{\max}(t) = 0\ , \end{cases}$$

and let r_a be the standard reward for picking up an apple. We add a small bias $b > 0$ so there is always some benefit to picking more apples. The shaped reward when an apple is picked by an agent is then

$$r_i^{\text{ineq}}(t) = \begin{cases} \big(1 - \tilde{c}_i(t)\big) + b & \text{if } c_{\max}(t) > 0, \\ 1 & \text{if } c_{\max}(t) = 0\ . \end{cases}$$

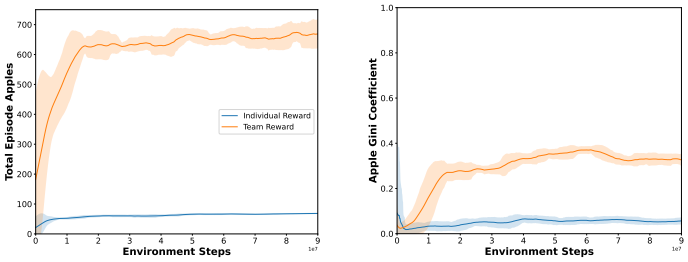
Thus, the current leader (with $\tilde{c}_i = 1$) receives b for additional apples (we use $b \approx 0.1$), which provides a small incentive to raise the maximum while still making it more valuable in the long run to clean so others can catch up. Agents with fewer apples than the leader earn a larger fraction of the apple reward.

5 Experiments

We compared the joint team policies learnt (in terms of total output and inequality) under four different reward functions — team reward, individual reward, fractional rewards and diminishing rewards.

5.1 Independent vs. team rewards

Our hypothesis is two-fold: (i) a team reward will increase total productivity, and (ii) it will induce specialisation—some agents harvest while others clean—which creates inequality in apples collected across agents. To test this, we compare



(a) Total Team Apples (b) Gini Coefficient

Fig. 3: Gini coefficient and total team productivity with respect to apples collected by the agents over the course of training. The Gini coefficient is a measure of inequality, where 0 is perfect equality and 1 is perfect inequality.

agents trained with independent rewards against agents trained with a shared team reward. Figure 3a shows total apples per episode over training. It is clear to see that under the team reward, agents can collect many more apples per episode than under the independent reward. However, there is an uneven distribution of apples collected by the agents. To quantify the inequality, we compute the Gini coefficient of the apples collected by the agents and plot the result in Figure 3b alongside the total productivity of the team.

5.2 Fractional rewards

To compare the effectiveness of the fractional reward, we plot the total number of apples collected by the team against that of the team trained using a fully independent reward. We tested different team reward weights α. Finally, to see the effect it had on inequality, we plot the Gini coefficients during training and the results are shown in Figure 4.

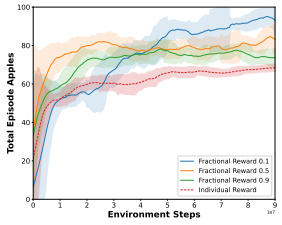
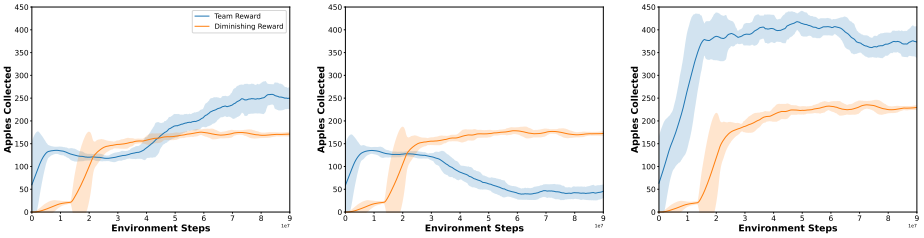


Fig. 4: Fractional team rewards with varying team reward weight α.

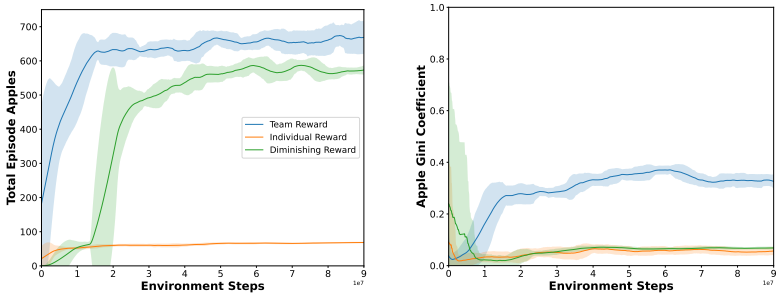
5.3 Inequality aware diminishing returns

We once again plot the per-agent number of apples collected over the course of training. In Figure 5, one can see that initially the agents learn to pick apples greedily, like in the independent reward case. However, after a brief plateau, the agents transition to collecting significantly more apples. Suggesting that they have learnt the benefit of cleaning the river and have learnt to share the apples that are grown. When we once again plot the Gini coefficient and total productivity over time, we see that the team's productivity approaches that of



(a) Agent 0 (b) Agent 1 (c) Agent 2

Fig. 5: Per agent number of apples collected over the course of training. The agents learn to pick apples greedily, like in the independent reward case. However, after a brief plateau, the agents then suddenly start to collect significantly more apples. Suggesting that they have learnt the benefit of cleaning the river and have learnt to share the apples that are grown.



(a) Total Team Apples (b) Gini Coefficient

Fig. 6: Gini coefficient of the apples collected by the agents over the course of training. The Gini coefficient is a measure of inequality, where 0 is perfect equality and 1 is perfect inequality.

the fully cooperative team, while simultaneously the inequality remains at levels similar to that of the independent reward case. This is a clear indication that the agents have learnt to share the apples that are grown and have learnt to clean the river. Moreover, it is to be expected that the total productivity will be slightly lower than the fully cooperative reward because a price is paid for not letting the agents specialise (i.e. some clean and some pick). However, the gains in terms of equality and fairness may make the sacrifice in productivity worthwhile, depending on the setting and the objectives.

5.4 Summary of findings

Finally, to summarise our findings for all of the different reward settings, we provide a table with different measures of the performance of the agents. In

Table 1: Summary of performance across reward schemes. Outcome in terms of total apples collected, utilitarian welfare, egalitarian welfare, and the Gini coefficient.

Reward	Total	Utilitarian	Egalitarian	Gini
Independent	67.3	22.4	20.6	0.05
Shared	723.4	241.1	116.1	0.33
Fractional	90.6	30.2	24.9	0.11
Diminishing	573	191.2	173.8	0.07

particular, we consider the total apples collected by the agents, the Gini coefficient of the apples collected by the agents, the utilitarian welfare and the egalitarian welfare. We can see from the table that the fractional team rewards only marginally increased team productivity, but also increased the inequality slightly. The Egalitarian welfare also only went up slightly. Diminishing returns, on the other hand, significantly improved total team yield over independent rewards, nearly achieving the levels of the shared reward. Moreover, the Egaleterian welfare of the diminishing returns was actually higher than for the shared reward. Finally, the inequality only went up slightly and was still significantly better than the shared rewards.

6 Conclusion

In this work, we investigate how welfare and inequality considerations can be used to improve the policy of agents in a MARL setting, such that the society remains reasonably socially balanced, while overall performance is not badly affected. We studied how reward design can steer prosocial behaviour in a stylised multi-agent social dilemma. In an apple-harvesting environment that balances private gains (harvesting) with a public good (cleaning), we compared independent rewards, fully shared team rewards, fractional team rewards, and an inequality-aware diminishing-returns shaping. Independent rewards harmed long-run productivity via over-harvesting; shared rewards delivered high output but high inequality; fractional spillovers were too weak. Inequality-aware diminishing returns nearly matched the shared-reward output while keeping inequality low, with agents shifting from early greedy harvesting to later cleaning.

This pilot study on balancing productivity and equality/fairness in MARL indicates some important results for both MARL, but potentially beyond, to true social interactions between humans and between humans and agents. The environment and models are deliberately simple, focusing on on-policy learning, shaping with global statistics and with intuitive metrics. Noteably, we have not studied generalisation, or application to more complex settings, in order to highlight the simplicity of the approach. It is clear that there are other avenues that

will be important to explore in the future, including ethical trade-offs around specialisation. Despite these constraints, we believe that this study highlights inequality as a controllable dimension alongside aggregate performance.

Future work should generalise and test diminishing returns across diverse social dilemmas, learn shaping functions from data or welfare objectives, combine with centralised critics or communication, and study robustness and incentive compatibility. A key direction is evaluating large LLM agents in analogous settings, comparing prompt-only steering with training under diminishing-returns rewards to reduce over-optimisation for individual benefit. Overall, simple reward shaping appears promising for retaining most efficiency gains of cooperation while materially reducing inequality.

Acknowledgments. We would like to thank the Cooperative AI Foundation for providing the neccessary compute for this project.

Disclosure of Interests. None to declare.

Bibliography

[1] Albrecht, S.V., Christianos, F., Schäfer, L.: Multi-agent reinforcement learning: Foundations and modern approaches. MIT Press (2024)

[2] Aloor, J.J., Nayak, S.N., Dolan, S., Balakrishnan, H.: Cooperation and fairness in multi-agent reinforcement learning. ACM Journal on Autonomous Transportation Systems **2**(2) (2024)

[3] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)

[4] Bellman, R., Kalaba, R.: Dynamic programming and statistical communication theory. Proceedings of the National Academy of Sciences **43**(8), 749–751 (1957)

[5] Chevaleyre, Y., Dunne, P.E., Endriss, U., Lang, J., Lemaitre, M., Maudet, N., Padget, J., Phelps, S., Rodrígues-Aguilar, J.A., Sousa, P.: Issues in multiagent resource allocation (2005)

[6] Claus, C., Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems. AAAI/IAAI **1998**, 2 (1998)

[7] Cusumano-Towner, M., Hafner, D., Hertzberg, A., Huval, B., Petrenko, A., Vinitsky, E., Wijmans, E., Killian, T., Bowers, S., Sener, O., et al.: Robust autonomy emerges from self-play (2025)

[8] De Witt, C.S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P.H., Sun, M., Whiteson, S.: Is independent learning all you need in the starcraft multi-agent challenge? (2020)

[9] Devlin, S., Yliniemi, L., Kudenko, D., Tumer, K.: Potential-based difference rewards for multiagent reinforcement learning. In: AAMAS. p. 165–172. IAAMAS (2014)

[10] Duggan, J.: Avoiding the tragedy of the commons using reward shaping (May 2016)

[11] Dulac-Arnold, G., Levine, N., Mankowitz, D.J., Li, J., Paduraru, C., Gowal, S., Hester, T.: Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. Machine Learning **110**(9), 2419–2468 (2021)

[12] Foerster, J., Nardelli, N., Farquhar, G., Afouras, T., Torr, P.H., Kohli, P., Whiteson, S.: Stabilising experience replay for deep multi-agent reinforcement learning. In: International conference on machine learning. PMLR (2017)

[13] Foerster, J.N., Chen, R.Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., Mordatch, I.: Learning with opponent-learning awareness (2018)

[14] Formanek, C., Tilbury, C.R., Shock, J.P.: Opportunities of reinforcement learning in south africa's just transition (2024)

[15] Grupen, L., Selman, B., Lee, G.K.: Cooperative multi-agent fairness and equivariant policies. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36 (2022)

[16] Guo, Z., Shi, S., Willis, R., Tomilin, T., Leibo, J.Z., Du, Y.: Socialjax: An evaluation suite for multi-agent reinforcement learning in sequential social dilemmas (2025)
[17] Gupta, J.K., Egorov, M., Kochenderfer, M.: Cooperative multi-agent control using deep reinforcement learning. In: International conference on autonomous agents and multiagent systems. Springer (2017)
[18] Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., et al.: Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143 (2025)
[19] Han, S., Zhang, Q., Yao, Y., Jin, W., Xu, Z.: Llm multi-agent systems: Challenges and open problems. arXiv preprint arXiv:2402.03578 (2024)
[20] Hansen, E.A., Bernstein, D.S., Zilberstein, S.: Dynamic programming for partially observable stochastic games. In: AAAI (2004)
[21] Howard, R.A.: Dynamic programming and markov processes. MIT Press (1960)
[22] Hughes, E., Leibo, J.Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al.: Inequity aversion improves cooperation in intertemporal social dilemmas. Advances in neural information processing systems **31** (2018)
[23] Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J.Z., De Freitas, N.: Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: International conference on machine learning. pp. 3040–3049. PMLR (2019)
[24] Jiang, J., Lu, Z.: Learning fairness in multi-agent systems. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
[25] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
[26] Leibo, J., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T.: Multi-agent reinforcement learning in sequential social dilemmas. In: AAMAS. ACM (2017)
[27] Lupu, A., Precup, D.: Gifting in multi-agent reinforcement learning. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. pp. 789–797 (2020)
[28] Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021)
[29] Marshall, A.: Principles of economics. Springer (2013)
[30] Moulin, H.: Fair division and collective welfare. MIT press (2004)
[31] Nash, J.F.: Non-cooperative games. In: The Foundations of Price Theory Vol 4, pp. 329–340. Routledge (2024)
[32] Ng, A.Y., Harada, D., Russell, S.: Policy invariance under reward transformations: Theory and application to reward shaping. In: Icml. vol. 99, pp. 278–287. Citeseer (1999)

[33] Nguyen, D.T., Kumar, A., Lau, H.C.: Credit assignment for collective multiagent rl with global rewards. Advances in neural information processing systems **31** (2018)
[34] Perolat, J., Leibo, J.Z., Zambaldi, V., Beattie, C., Tuyls, K., Graepel, T.: A multi-agent reinforcement learning model of common-pool resource appropriation. Advances in neural information processing systems **30** (2017)
[35] Pretorius, A., Cameron, S., Van Biljon, E., Makkink, T., Mawjee, S., du Plessis, J., Shock, J., Laterre, A., Beguir, K.: A game-theoretic analysis of networked system control for common-pool resource management using multi-agent reinforcement learning. Advances in neural information processing systems **33**, 9983–9994 (2020)
[36] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
[37] Rutherford, A., Ellis, B., Gallici, M., Cook, J., Lupu, A., Ingvarsson Juto, G., Willi, T., Hammond, R., Khan, A., Schroeder de Witt, C., et al.: Jaxmarl: Multi-agent rl environments and algorithms in jax. Advances in Neural Information Processing Systems **37**, 50925–50951 (2024)
[38] Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation (2015)
[39] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms (2017)
[40] Sen, A.: On economic inequality. Oxford university press (1997)
[41] Sheikh, H.U., Bölöni, L.: Multi-agent reinforcement learning for problems with combined individual and team reward (2020)
[42] Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An introduction. MIT Press Cambridge (1998)
[43] Sutton, R.S., McAllester, D., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems **12** (1999)
[44] Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., Stone, P.: Deep reinforcement learning for robotics: A survey of real-world successes. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
[45] Wang, J., Zhang, Y., Kim, T.K., Gu, Y.: Shapley q-value: A local reward approach to solve global reward games. Proceedings of the AAAI Conference on Artificial Intelligence **34**(05), 7285–7292 (Apr 2020)
[46] Wolpert, D.H., Wheeler, K.R., Tumer, K.: Collective intelligence for control of distributed dynamical systems. Europhysics Letters (EPL) **49**(6), 708–714 (Mar 2000)
[47] Wolpert, D.H., Tumer, K.: Optimal payoff functions for members of collectives. Advances in Complex Systems **4**, 265–279 (2001)
[48] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., Wu, Y.: The surprising effectiveness of ppo in cooperative multi-agent games. Advances in neural information processing systems (2022)

[49] Zhang, M., Vuong, R.D.C., Luo, H.: No-regret learning for fair multi-agent social welfare optimization (2024)
[50] Zheng, S., Trott, A., Srinivasa, S., Parkes, D.C., Socher, R.: The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science advances **8**(18), eabk2607 (2022)
[51] Zhou, J.L., Hong, W., Kao, J.C.: Reciprocal reward influence encourages cooperation from self-interested agents (2025)
[52] Zimmer, M., Siddique, U., Weng, P.: Learning fair policies in decentralized cooperative multi-agent reinforcement learning. CoRR (2020)